Appendix

Appendix A1 Study characteristics: Beck Evaluation & Testing Associates, 2005 (randomized controlled trial)

Characteristic	Description
Study citation	Beck Evaluation & Testing Associates, Inc. (2005). *Progress in Mathematics © 2006: Grade 1 pre-post field test evaluation study.* New York: Sadlier-Oxford Division, William H. Sadlier, Inc.
Participants	The study included 186 first graders (96 students in the intervention group and 90 students in the comparison group) in eight classrooms across four schools. Within schools, classrooms were randomly assigned to the intervention or comparison group. For rating purposes, the sample for the analysis of the Terra Nova Mathematics Test included 186 students, and the sample for the analysis of the Terra Nova Math Computation Test included 181 students.
Setting	The eight classrooms were located in four elementary schools in four school districts in the eastern United States. Three of the schools were Catholic schools, and one was a public school. One pair of classrooms (one intervention and one comparison) was located in each of the participating schools.
Intervention	The intervention classrooms received the pre-publication version of *Progress in Mathematics © 2006* student edition materials, student workbooks, and teacher guides. The study indicated that those materials resembled as closely as possible the intended published version.
Comparison	The comparison classrooms used the 2000 version of *Progress in Mathematics.* This textbook series had been used in the participating schools for at least three years prior to the study. This intervention report regards *Progress in Mathematics © 2006* as a different math program from *Progress in Mathematics © 2000.* The WWC team compared the textbooks of both programs and found them to differ extensively in terms of content, assessment materials, organization, and presentation. Whereas the 2000 version emphasizes written computation skills, the © 2006 version focuses on mathematical language and problem solving in addition to computation. Information received from the developer confirmed this difference between programs.
Primary outcomes and measurement	Students were tested using the TerraNova Mathematics and Math Computation Tests (see Appendix A2 for more detailed descriptions of outcome measures).[1]
Teacher training	Intervention group teachers received a pre-implementation orientation from the developer's editorial staff. They also received ongoing editorial department support through in-person visits and by phone throughout the study.[2] The comparison group teachers already had previous training and experience with their current textbooks.

1. The study reported on student outcomes using an additional outcome measure, the Custom Test, which did not meet WWC evidence screens because of differential attrition of students in the intervention and comparison groups.
2. This study does not provide information about whether this level of training and ongoing support is reflective of the program's typical implementation.

Appendix A2 Outcome measures in the math achievement domain

Outcome measure	Description
TerraNova Mathematics Test	Level 11, Form C of the CAT TerraNova series, second edition (McGraw-Hill, 2001; as cited in Beck Evaluation and Testing Associates, 2005) is a standardized nationally normed test. This mathematics test includes 47 items and is part of the CAT Basic Battery.
TerraNova Mathematics Computation Test	Level 11, Form C of the CAT TerraNova series, second edition (McGraw-Hill, 2001; as cited in Beck Evaluation and Testing Associates, 2005) is a standardized nationally normed test. This test includes 20 mathematics computation items and is part of the CAT Plus portion of the Terra Nova series.

Appendix A3 Summary of study findings included in the rating for the math achievement domain[1]

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Progress in Mathematics* group[3]	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*Progress in Mathematics* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Beck Evaluation & Testing Associates, 2005 (randomized controlled trial)[8]								
TerraNova Mathematics Test	Grade 1	8/186	40.62 (4.30)	37.70 (5.80)	2.92	0.57	ns	+22
TerraNova Mathematics Computation Test	Grade 1	8/181	15.50 (2.70)	16.80 (3.30)	–1.30	–0.43	ns	–17
Domain average[9] for math achievement (Beck Evaluation & Testing Associates, 2005)						0.07	ns	+3

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. Subtest findings from the same study are not included in these ratings, but are reported in Appendix A4.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The intervention group mean equals the comparison group mean plus the mean difference.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The mean difference and effect size were calculated using the difference in difference approach, which takes baseline student scores into account.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between the groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Beck Evaluations & Testing Associates (2005), corrections for clustering and multiple comparisons were needed, so the significance levels differ from those reported in the original study.
9. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A4 *Progress in Mathematics © 2006* rating for the math achievement domain

The WWC rates an intervention's effects in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of math achievement, the WWC rated *Progress in Mathematics © 2006* as having no discernible effects. It did not meet the criteria for other ratings (positive effects, potentially positive effects, mixed effects, potentially negative effects, and negative effects) because the single study that met WWC standards did not show statistically significant or substantively important effects.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: None of the studies shows a statistically significant or substantively important effect, either *positive* or *negative.*

  Met. The single study that assessed outcomes in this domain showed indeterminate effects.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. No studies showed statistically significant or substantively important positive effects.

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. No studies showed statistically significant or substantively important negative effects.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed a statistically significant or substantively important positive effect.

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. The single study on *Progress in Mathematics © 2006* showed indeterminate effects.

Mixed effects: Evidence of inconsistent effects as demonstrated through either of the following criteria.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, and at least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed a statistically significant or substantively important effect.

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an *indeterminate* effect than showing a statistically significant or substantively important effect.

  Not met. No studies showed a statistically significant or substantively important effect.

(continued)

Appendix A4 *Progress in Mathematics © 2006* rating for the math achievement domain *(continued)*

Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence

- Criterion 1: At least one study showing a statistically significant or substantively important *negative* effect.

  Not met. No studies showed a statistically significant or substantively important negative effect.

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

  Met. No studies showed a statistically significant or substantively important positive effect.

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *negative* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. No studies showed a statistically significant negative effect.

- Criterion 2: No studies showing statistically significant or substantively important *positive* effects.

  Met. No studies showed a statistically significant or substantively important positive effect.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A5 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
Math achievement	1	4	181	Small

1. A rating of "moderate to large" requires at least two studies and two schools across studies in one domain, and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is "small."